Key Qualities of Conversational Recommender Systems: From Users' Perspective

Yucheng Jin
Department of Computer Science, Hong Kong Baptist University
Hong Kong, China
yuchengjin@comp.hkbu.edu.hk

Li Chen
Department of Computer Science, Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Wanling Cai
Department of Computer Science, Hong Kong Baptist University
Hong Kong, China
cswlcai@comp.hkbu.edu.hk

Pearl Pu
School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT

An increasing number of recommender systems enable conversational interaction to enhance the system's overall user experience (UX). However, it is unclear what qualities of a conversational recommender system (CRS) are essential to determine the success of a CRS. This paper presents a model to capture the key qualities of conversational recommender systems and their related user experience aspects. Our model incorporates the characteristics of conversations (such as adaptability, understanding, response quality, rapport, humanness, etc.) in four major user experience dimensions of the recommender system: *User Perceived Qualities*, *User Belief*, *User Attitudes*, and *Behavioral Intentions*. Following the psychometric modeling method, we validate the combined metrics using the data collected from an online user study of a conversational music recommender system. The user study results 1) support the consistency, validity, and reliability of the model that identifies seven key qualities of a CRS; and 2) reveal how conversation constructs interact with recommendation constructs to influence the overall user experience of a CRS. We believe that the key qualities identified in the model help practitioners design and evaluate conversational recommender systems.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; • **Human-centered computing** → **User studies**; **Heuristic evaluations**.

KEYWORDS

Recommender systems, conversational recommender systems, user experience, questionnaire, user-centric evaluation.

HAI '21, November 9–11, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8620-3/21/11…$15.00
https://doi.org/10.1145/3472307.3484164

ACM Reference Format:
Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21), November 9–11, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3472307.3484164

1 INTRODUCTION

The new generation of conversational recommender systems (CRSs) enables users to interact with recommendations using natural human language. Traditionally, users click a rating button to tell the system if they like the recommended item or not, while in a CRS, users may say *"I like the melody of this song"* to express their preferences in more detail. The CRS's prominent feature is the natural interaction that facilitates user critiquing [1] and user exploration [2] in recommender systems.

The implementation of the CRS relies on two major AI technologies, i.e., recommendation technique and natural language processing technique. However, current evaluation frameworks of recommender systems mainly focus on recommendations but ignore conversations, which might not be sufficient to assess the ultimate success of a CRS. Previous studies [3–5] have shown the limitations of only considering objective metrics in evaluating recommender systems. We have seen several user-centric evaluation frameworks, such as the widely used ResQue questionnaire [6] and Knijnenburg et al. proposed framework [7], to measure user experience of recommendations. The user-centric evaluation of a CRS can be even more important and challenging. This is because a CRS intends to improve the overall user experience (UX) of recommendations through a more natural human-computer interaction. Recent studies [1, 2, 8] on CRSs have considered some prominent qualities of both recommendations and conversations to gauge the user experience. However, it is unknown how these qualities interact to influence user behavioral intentions and which qualities are more crucial to the success of a CRS. To this end, we introduce a model that investigates how the qualities of a CRS influence behavioral intention to use the CRS. We develop this model based on existing user-centered evaluation work on recommender systems and conversational agents. The structure of our model follows the four key dimensions: User Perceived Qualities, User Belief, User

Attitudes, and Behavioral Intentions, which are identified in *ResQue*, a widely used evaluation framework for recommender systems [6].

The development of our model follows an empirical research methodology. We validate the new model by conducting a user study with a conversational music recommender. The results of our study confirmed that our model achieved sufficient (high) reliability for assessing user experience of a CRS and the hypothesized paths in the model have been validated. With our **CRS-UX** model, we attempt to answer the following research questions:

- **RQ1**: *Which qualities of a CRS are particularly important in terms of their influence on intention to use the CRS?*
- **RQ2**: *How do conversation constructs interact with recommendation constructs to influence user experience of a CRS?*

The main contributions of our work are two-fold:

(1) We develop a unified model that identifies key qualities of conversational recommender systems from the users' perspective. In particular, the new model reveals how the qualities of conversation correlate with the qualities of recommendation and how they altogether can be used to design and evaluate a CRS.
(2) We employ an empirical research method to validate the model by conducting a user study (N=173) with a conversational music recommender.

We structure the paper as follows. We first review the related work on conversational recommender systems and existing user-centric evaluation frameworks for recommender systems and conversational agents respectively. After that, we explain the development process of our model including the constructs and the hypothesized relations. We then validate the model by presenting a user study including experimental setup and data analysis. Finally, we discuss the model and study results, and conclude with the limitations and future plans of our work.

2 RELATED WORK

2.1 Conversational Recommender Systems

The conversational recommender system (CRS) allows users to find their interesting recommendations with multi-turn interactions [9]. Unlike traditional recommender systems that only support a one-shot interaction, i.e., presenting one or a list of recommended items based on users' past behavior [10], the CRS can interactively elicit users' current preferences from their feedback and build a more comprehensive user model to make better recommendations [11]. According to a recent survey on CRS [9], some earlier CRSs were built on the graphical user interface (GUI), such as critiquing-based systems [12] where users give feedback on recommendations by choosing some pre-defined critiques. However, the recent advance of natural language technology has led to an increased interest in building a CRS on the *conversational user interface (CUI)*, where users interact with the recommender system through conversation [13, 14].

However, most researchers evaluated CRS techniques using offline experiments, which usually simulated user behavior, for example, answering preference-related questions or giving feedback on recommendations, based on their historical data [15]. With simulated data, they separately measured the recommendation performance by adopting accuracy measures (for example, Average Precision, RMSE, and Recall) [11], and/or assessed the system's responses using linguistic measures like BLEU score [16], but this may not reflect the overall quality of the system. Such simulated evaluation did not consider that users may develop new preferences after they explore recommendations during the conversation. Thus it may not inform us about the evaluation results in real-world situations. In contrast, empirical studies carried with user-centric evaluation can better measure the system's effectiveness in realistic situations. It typically requires participants to use the system to complete a specific task (for example, finding music for a party) and then assesses their quality perception of the system [1, 2]. But to the best of our knowledge, so far, rare work has identified key qualities of a CRS from the users' perspective. To this end, our work aims at developing a model to fill in this vacancy.

2.2 UX Metrics for Recommender Systems

Given the limitations of evaluation methods based on objective metrics, several studies proposed different UX metrics for recommender systems. The most influential ones are *ResQue* [6] and the framework proposed by Knijnenburg et al [7]. *ResQue* is a unifying evaluation framework that measures the qualities of the recommended items and analyzes how these qualities influence Behavioral Intentions through User Beliefs and Attitudes. Knijnenburg et al. proposed a framework to explain users' behavior through a set of constructs organized in a structure relating the objective system aspects, subjective system aspects (i.e., the perceived qualities of the system), experience constructs (i.e., how users experience the system), personal characteristics, and situational characteristics. Numerous studies have employed one of the two frameworks to evaluate various types of recommender systems ranging from social learning recommendations [17], music recommendations [1], movie recommendations [18], to product recommendations [19]. In addition, we find some evaluation questionnaires that focus on some specific UX constructs of recommender systems such as explanation [20], trust [21], inspectability, and user control [22]. As we mainly focus on user perception of recommendations, we choose ResQue in our study.

2.3 UX Metrics of Conversational Agents

From a technical point of view, the evaluation metrics of conversational agents (CA) have identified several key components, such as the performance of natural language understanding (NLU) component [23] and natural language generation (NLG) component [24, 25]. One popularly used evaluation framework for CA is PARADISE [26], which mainly focuses on assessing task success rate and dialogue cost (for example, dialogue time, number of utterances, agent response delay). Given that CA usability can significantly influence the demonstration and perception of CA functionality [27], we especially review the metrics of measuring the quality of conversational experience. Walker et al. [26] proposed a general performance model of system usability for spoken dialogue agents, which includes a subjective metric of user satisfaction and three objective metrics of dialogue efficiency, dialogue quality, and task success. Ruttkay et al. [28] proposed a

framework for comparing and evaluating embodied conversational agents and identified the general and most important issues of evaluating CA. Kuligowska's proposed metrics include performance, usability, and overall quality of commercial conversational applications [29]. Based on the quality attributes of chatbot development and implementation, Radziwill and Benton [30] proposed a quality assessment method and introduced an analytic hierarchy process (AHP) for quality metrics selection. Guerini et al. [31] provided a novel methodology to assess the impact of the agent's interaction strategies on the quality of experience. The metrics mainly consider two dimensions: Quality of Service (QoS) and Quality of Experience (QoE). Zhao et al. [32] developed a metric based on the theories of negotiation and communication. It characterizes the interaction into dimensions of rapport, such as positivity, attentiveness, and coordination. PEACE model [33] identifies four essential qualities of a chatbot (including politeness, entertainment, attentive curiosity, and empathy) that influence users' intention to use open-domain chatbots.

Each metric's constructs vary, while some constructs are common in some metrics such as task ease, performance, and satisfaction. Some of their constructs [29, 32] consider features more about communications such as language skill, coordination, rapport, while some metrics include more comprehensive constructs such as future use [34], affect [30], trust [28], and ease of use [31]. Besides, four of above mentioned metrics specialize in particular types of CA, such as commercial CA, task-oriented CA, social negotiation CA, and open-domain CA. Similar to the evaluation of recommendations, objective measures are insufficient to gauge the effectiveness and user experience of a conversational recommender system (CRS). Since a CRS is a task-oriented conversational agent, the evaluation should consider the success of dialogue if the agent helps users find the recommended items of their interests.

3 MODEL DEVELOPMENT

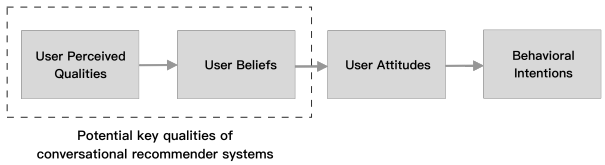

Figure 1: Abstract Levels of Key Qualities in CRS-UX.

We develop our model **CRS-UX** based on *ResQue* model [6], which consists of four dimensions: *Perceived System Qualities*, *User Beliefs*, *User Attitudes*, and *Behavioral Intentions*. Each dimension contains several constructs carefully derived from prior work related to the user experience of recommender systems. Our model considers the constructs of *Perceived System Qualities* and *User Beliefs* as **potential qualities of a CRS**, the constructs of *User Beliefs* measure a higher level of user perceived quality of the system. We, therefore, determine the constructs that influence Behavioral Intentions as the **key qualities** of conversational recommender systems in our model (see Figure 1). We identify the potential key qualities of conversational agents from our surveyed papers. We first propose initial sample questions and then conduct a pilot study to refine the constructs and their contained questions. Our model organizes the question items into the four dimensions to clearly describe how these potential qualities of a CRS influence User Attitudes, and furthermore Behavioral Intentions. After modifying and dropping the redundant and confusing questions, the final model contains nineteen constructs and thirty-seven questions (see Table 2). The following sections explain what the constructs are supposed to measure and review the relevant studies that have inspired our model's development.

3.1 Perceived System Qualities

Perceived system qualities (PSQ) are defined as the first layer of **CRS-UX**, in which we mainly measure how users perceive the major characteristics of the recommender system such as recommendation accuracy, interaction adequacy, and those of conversational agents including positivity, attentiveness, coordination, understanding, adaptability, and response quality. We omit several constructs of the original *ResQue* model due to the unique feature of CRSs. For example, we omit *diversity* because it is a measure for a set of recommended items rather than a single item usually recommended by a CRS at a time, and we also exclude the construct of *interface adequacy* as it focuses on the graphical user interface. Our study validates some existing effects of User Perceived Qualities of recommendations on User Beliefs, Attitudes, and Behavioral Intentions found in the context of using conversational interaction. Moreover, we investigate how conversations constructs correlate with recommendations constructs to influence the other three dimensions.

Recommendation Accuracy. Perceived accuracy measures to what extend users feel the recommendation matching their interests and preferences. It can compensate for the limitation of objective accuracy [35], to indicate how good the recommendation could be from the users' perspective.

Explanation. This construct measures the system's ability to explain its recommendations. Explainable recommender systems tend to improve the trustworthiness and transparency of the system [36]. Several works [37, 38] have proposed different approaches to design and evaluate explanations of recommender systems. Explanations affect a user's mental model of the recommender system [37], however they may also negatively influence perceived recommendation accuracy [39] .

Novelty. Novelty is one of the most discussed beyond-accuracy metrics for recommender systems, which gauges the extent to which the recommendation is new or unknown to users. Novelty is particularly important to a recommender system that aims to support user exploration and discovery of new items. Novelty is always discussed together with "serendipity"; however, Herlocker [40] argued that recommendation of high serendipity should be not only new but also surprising. Despite the nuances of the two words, we do not distinguish them in our user study to avoid users' confusion. Novelty is usually positively correlated with some of other metrics like diversity and coverage [41].

Interaction Adequacy. This construct mainly measures the system's ability to elicit and refine user preferences through user interaction. However, some recommender systems may implicitly adapt to user preferences based on their interaction behaviors. Despite

more interaction efforts, a CRS tends to improve user experience through dialogue-based conversations [42]. Unlike the single-shot elicitation model in traditional RS, preference elicitation is usually an incremental process in a CRS [43]. Similar to the common interaction strategies of critiquing based recommender systems [44], a CRS allows users to give feedback by rating items or specifying the attributes of their preferred items.

CUI Positivity. The "Rapport" of conversations consists of three essential components[1] including Positivity, Attentiveness, Coordination [45]. Positivity is the first component of Rapport and it corresponds to the perceived mutual friendliness and caring in the communication. For example, positivity may determine the tone and vocabulary of conversations [45].

CUI Attentiveness. Attentiveness is the second component of Rapport. It measures if a system establishes a focused and cohesive interaction by expressing mutual attention and involving each other. CUI attentiveness closely relates to the other two components, namely Positivity and Coordination [45].

CUI Coordination. Coordination is the final component of Rapport that examines if the communication is synchronous and harmonious [45]. Coordination is more critical to Rapport than the other two components in the late communication phase. Besides, coordination tends to arouse empathy; thus, communicators respect each other.

CUI Understanding. Understanding is the key performance indicator of conversational agents, which measures an agent's ability to understand users' intents. The evaluation of the natural language understanding (NLU) module of a CA usually measures its performance of classifying intents and extracting entities, and its confidence scores stability [46]. In our work, we aim to measure user perceived understanding of a CRS.

CUI Adaptability. Adaptability measures a system's ability to adapt to users' behavior and preferences during the conversation. The adaptability is usually associated with personalization, i.e., whether a system can personalize its replies by adapting to the user's emotions and historical behavior [47]. Other adaptive agents can learn users' vocabularies to engage with community members [48], and adjust the length of conversation according to the context [49]. This construct in our model particularly assesses if the system adapts to the user's preference for items.

CUI Response Quality. We measure the response quality of a CRS from two aspects: content quality (informativeness) and the pace of interaction (fluency), which have been widely adopted for evaluating the quality of conversational agents' responses [50]. According to an evaluation framework for CA [51], informativeness and fluency are human judgments of conversational aspects, which in turn influence human judgments of the overall quality of the conversations such as engagingness and humanness.

3.2 User Beliefs

The constructs of User Beliefs in our **CRS-UX** model measure a higher level of user perception of a system, which are influenced by the constructs of Perceived Qualities. We concretely incorporate three constructs related to user perception of conversations in this evaluation layer, which are CUI rapport, CUI Engagingness, and CUI Humanness, in addition to the constructs in the original *ResQue* model (for example, User Control, Perceived Usefulness, Perceived Ease of Use, etc.) [6]. Overall, this layer's constructs focus on the effectiveness of a CRS in supporting users to perform specific tasks, such as decision making and exploration of new items.

User Control. User control measures the level of controllability users perceive while interacting with the recommender. Previous studies show the positive effects of user control on multiple user experience factors such as perceived qualities [52] and overall user satisfaction [53]. To address the challenges of designing personalized user control mechanisms for recommender systems [54], several studies [55, 56] suggest different user control mechanisms that are tailored to some personal characteristics such as domain knowledge, trusting propensity, and persistence.

Perceived Usefulness. Perceived usefulness measures the competence of the system in terms of supporting users in performing tasks [21]. Users may judge the usefulness of a recommender by comparing it with their experience of performing a similar task without the recommender's support. It was found that the perceived usefulness influences the users' willingness to share their data for improving E-commerce recommendations [57]. In our model, perceived usefulness particularly measures the extent to which the system supports decision making.

Perceived Ease of Use. Perceived ease of use can be measured physically and mentally. The psychical measures include completion time of performing a task and learning curve of using a new system. Regarding mental measures, many user studies of recommender systems employ the NASA-TLS evaluation framework [58] to assess users' cognitive load. We believe both physical efforts and cognitive load will influence the perceived ease of use of a CRS. Similar to *ResQue* [6], we use the subjective questions to measure this construct.

Transparency. Transparency of a system enables users to understand the inner logic of the recommendation process. Transparency closely relates with user control and explanation, the implementation of which constructs tends to positively influence users' perceived accuracy [59], intention to buy [60], and overall satisfaction [19]. Intuitively, transparency is also supposed to influence user trust positively [61]. However, a user study of content-based art recommender does not show such an effect [39]. Besides, Kizilcec [62] argues that designers should find a proper degree of interface transparency for building trust, as too much transparency may impair user trust.

CUI Rapport. It measures if users perceive a rapport while communicating with the conversational agent. Several studies investigate different approaches to help agents develop and maintain a communication rapport with users. For instance, Novick and Gris [63] suggest increasing amplitude of nonverbal behaviors to establish a rapport; and Riek et al. [64] enable human-robot rapport via real-time head gesture mimicry.

CUI Engagingness. In a broader sense, engagingness refers to the quality of the user experience that emphasizes user desire to continuously use a product for a long time [65]. It can be influenced by many factors such as positive affect, aesthetic and sensory appeal, novelty, and perceived user control [66]. See et al. [51] define engagingness as an overall quality measure of conversation.

[1]These three components closely correlate, and each component's relative importance may change over communication time.

CUI Humanness. Humanness is also an overall quality measure of conversation, as it gauges the extent to which an agent behaves like a human. Adiwardana et al. [67] propose user-centric evaluation metrics, i.e., Sensibleness and Specificity Average (SSA), which capture key elements of a human-like multi-turn conversation. Moreover, many studies show various factors that may influence user perception of humanness, such as anthropomorphic visual cues [68], the presence of typos and capitalized words in the responses [69], typeface [70], and conversational skills [71]. However, a study suggests avoiding small talk and maintaining a formal tone to reduce humanness in a service-oriented context [72].

3.3 User Attitudes

User Attitudes assess users' overall feelings towards a conversational recommender system. Compared with the constructs of User Beliefs, the constructs of Attitudes are less likely to be influenced by the short-term experience of using the system. The typical constructs of Attitudes include user trust, user confidence, and overall satisfaction.

Trust. Trust significantly influences the overall success of recommendations. Incorporating the concept of trust into a collaborative filtering framework tends to increase the predictive accuracy of recommendations [73, 74]. Kunkel et al. [75] suggest that recommenders should provide richer explanations to increase a system's trustworthiness. Pu and Chen [21] explore the potential of building users' trust with explanation interfaces for recommender systems. Besides, system reputation [76], tasks and contexts [77], cultural differences [78], and familiarity [79] may also influence user trust. For a CRS, the trust may be related to recommendations, conversations, or both. Although Przegalinska et al. [80] propose a new methodology to measure chatbot performance based on user trust, the trust in conversations does not have a unified definition and measurement yet [81]. Since the design of human-AI conversations often refers to human-human interactions, we need to consider some constructs of social communication if focusing on the conversations. Therefore, in our current model, the trust primarily emphasizes recommendations.

Confidence. Confidence indicates if the system can convince users of recommended items. In other words, it measures how much the user is confident in accepting the recommendation. Hoxmeier et al. [82] investigate the effects of gender and technical experience on user confidence in electronic communication. For decision support systems, the level of presenting uncertainty information can influence user confidence in decision making [83]. *Overall Satisfaction* This construct is an overall measure of users' attitudes and opinions towards a conversational recommender system. It allows subjects to provide general feedback to the whole system. Several studies show increased user satisfaction by integrating user personality traits [84] and domain knowledge [85] into the process of generating recommendations. Besides, a large-scale user study [86] shows a significantly positive effect of recommendation serendipity on user satisfaction.

3.4 Behavioral Intentions

Behavioral Intentions towards a system are related to user loyalty, which measure the likelihood users are willing to use the system in the future, accept/purchase resulting recommendations, and recommend the system to their acquaintances. Wang et al. [87] propose a research model to investigate four factors that influence Behavioral Intentions of using a recommender, which includes performance expectancy, effort expectancy, social influence, and trust. Besides, Shin [88] shows the direct effects of trust and satisfaction on intention to use.

Table 1: Demographics of 173 Participants

	Item	Frequency	Percentage (%)
Age	19-25	80	46.24%
	26-30	35	20.23%
	31-35	19	10.98%
	36-40	13	7.51%
	>40	26	15.03%
Gender	Male	90	52.02%
	Female	80	46.24%
	Other	3	1.73%

4 MODEL VALIDATION

To validate our proposed model, we follow a psychometric methodology to test the model's internal reliability and convergent validity by performing confirmatory factor analysis (CFA) with the data collected from an empirical user study. Furthermore, we employ an advanced statistical model, structural equation model (SEM), to analyze the correlations among our model's constructs systematically.

4.1 Experimental Setup

We recruited subjects from Prolific,[2] a popularly used platform for academic surveys, to evaluate a research prototype of chatbot for music recommendations. More details about the prototype design and implementation are described in our prior work [2]. This study has been approved by our university's committee on the use of human & animal subjects in teaching and research. We then asked all subjects to find favorite songs by using this system. To ensure the quality of the experiment, we pre-screened users in Prolific using the following criteria: (1) participants should be fluent in English; (2) the number of the participant's previous submissions should be more than 100; (3) approval rate should be greater than 95%. The experiment took 20 minutes on average, and we compensated each participant £2.4. A total of 265 users participated in our study. We removed 38 participants' responses for extremely long duration in the study and 54 participants who failed to pass the attention check questions. [3] We finally kept the data of 173 participants, which meet the minimum sample size (N=100–150) for conducting SEM [89]. Table 1 presents demographics of those subjects.

[2] https://www.prolific.co/

[3] To ensure the quality of user responses, we set attention checking questions (for example, "Please indicate which number is an odd number?"). Besides, we checked if users' responses have certain patterns, for example, "AAAA", "ABAB", or showing conflicts to similar or reversing questions.

Table 2: Results of reliability test for latent factors. Constructs with single question items are included for completeness.

Constructs	Items	Internal Reliability		Convergent Validity		
		Cronbach alpha (0.5)	Item-total correlation (0.4)	Factor loading (0.5)	Composite reliability (0.8)	Variance extracted (0.5)
Perceived System Qualities						
1. Recommendation Accuracy	1					
The songs recommended to me match my interests.						
2. Explanation	1					
The music chatbot explains why the products are recommended to me.						
3. Novelty	4	0.9304			0.9329	0.7771
The music chatbot helps me discover new songs.			0.7581	0.7956		
The music chatbot provides me with surprising recommendations that helped me discover new music that I wouldn't have found elsewhere.			0.8986	0.9417		
The music chatbot provides me with recommendations that I had not considered in the first place but turned out to be a positive and surprising discovery.			0.8229	0.8639		
The music chatbot provides me with recommendations that were a pleasant surprise to me because I would not have discovered them somewhere else.			0.8748	0.9226		
4. Interaction Adequacy	3	0.7848			0.7952	0.5654
I find it easy to inform the music chatbot if I dislike/like the recommended song.			0.5855	0.7370		
The music chatbot allows me to tell what I like/dislike.			0.5723	0.6730		
I find it easy to tell the system what I like/dislike.			0.7333	0.8594		
5. CUI Attentiveness	1					
The music chatbot is interested in what I am saying.						
6. CUI Understanding	1					
The music chatbot understands what I say.						
7. CUI Adaptability	3	0.8117			0.8147	0.5955
I felt I am in sync with the music chatbot.			0.6393	0.7891		
The music chatbot adapts continuously to my preferences.			0.6388	0.7392		
I always have the feeling that this music chatbot learns my preferences.			0.7135	0.7826		
8. CUI Response Quality	4	0.8239			0.8325	0.5612
The music chatbot's responses are readable and fluent.			0.6111	0.6521		
Most of the chatbot's responses make sense.			0.6171	0.6817		
The pace of interaction with the music chatbot is appropriate.			0.7305	0.8377		
The music chatbot responds to my query/request quickly.			0.6525	0.7791		
User Beliefs						
1. User Control	1					
I feel in control of modifying my taste using this music chatbot.						
2. Perceived Usefulness	3	0.8180			0.8165	0.5976
The music chatbot helps me find the ideal item.			0.7000	0.7635		
Using the music chatbot to find what I like is easy.			0.6589	0.7951		
The music chatbot gives me good suggestions.			0.6578	0.7627		
3. Perceived Ease of Use	1					
I easily find the songs I was looking for.						
4. Transparency	1					
I understand why the songs are recommended to me.						
5. CUI Rapport	5	0.8947			0.8957	0.6337
The music chatbot is warm and caring.			0.7429	0.8111		
The music chatbot cares about me.			0.8148	0.8789		
I like and feel warm toward the music chatbot.			0.7899	0.8639		
I feel that I have no connection with the music chatbot.			0.6196	0.6465		
The music chatbot and I establish rapport.			0.7589	0.7895		
6. CUI Engagingness	1					
I feel it is entertaining and interesting to engage in a dialogue with this music chatbot.						
7. CUI Humanness	1					
The music chatbot behaves like a human.						
User Attitudes						
1. Trust	1					
This music chatbot can be trusted.						
2. Confidence	1					
I am confident I will like the items recommended to me.						
3. Overall Satisfaction	1					
Overall, I am satisfied with this music chatbot.						
Behavioral Intentions						
Intention to Use	3	0.9228			0.9249	0.8045
I will use this music chatbot again.			0.8407	0.9045		
I will use this music chatbot frequently.			0.8871	0.9437		
I will tell my friends about this music chatbot.			0.8068	0.8441		

4.1.1 Experimental Procedure. The procedure of the experiment contains the following steps: 1) Participants need to sign a consent form to accept General Data Protection Regulation (GDPR) before signing into our system with their Spotify accounts; 2)We ask participants to read a brief introduction about using the music CRS and fill a pre-study questionnaire; 3)They are allowed to try the system for 2 minutes. 4)They are asked to perform a task using the system, which is to find the top-5 favorite songs. 5)After finishing the task, we ask users to fill a post-study questionnaire according to **CRS-UX**. All the question items in the post-study questionnaire were measured on a seven-point Likert scale.

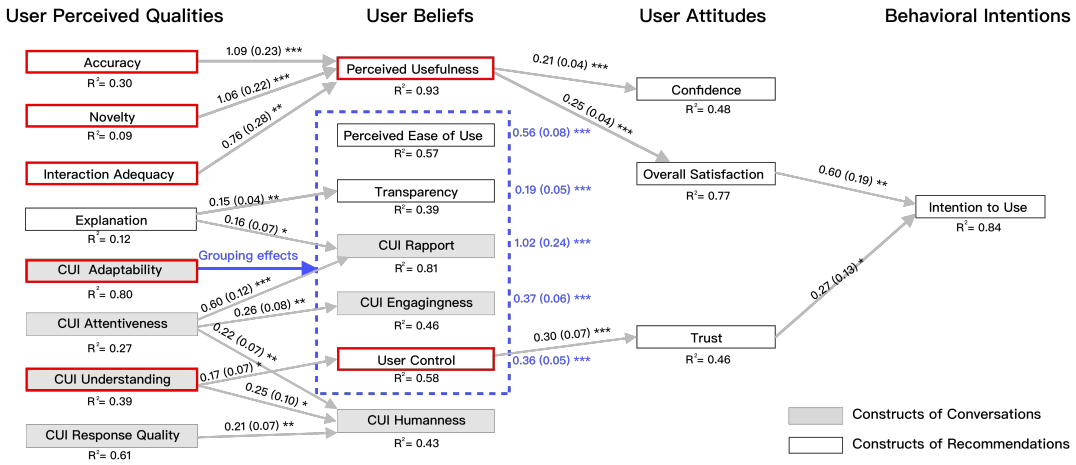

Figure 2: The structural equation modeling (SEM) results. Significance: *** $p < .001$, ** $p < .01$, * $p < .05$. R^2 is the proportion of variance explained by the model. Factors are scaled to have a standard deviation of 1.

4.2 Validity and Reliability of Model

We first validate each construct in our model and then perform path analysis using SEM. A confirmatory factor analysis (CFA) model consists of latent variables (or called factors), which cannot be measured directly and should be measured by at least three indicators [90]. By contrast, the observable variable is a variable that can be measured directly. Seven constructs of our model are latent variables, and the rest of the constructs are observable variables. Confirmatory Factor Analysis (CFA) aims to build both convergent and discriminant validity. Convergent validity ensures that a set of questions (indicators) measure the same latent factor. In contrast, discriminant validity ensures that the two latent factors' indicators do not measure the same factor. We iteratively adjust the model based on the factor loadings and correlation coefficient between two factors (for example, we may remove an indicator until the average variance extracted (AVE) of a factor is less than 0.4 [90], or we combine two factors if they strongly correlate). A latent variable should contain at least three indicators. Thus, if a latent variable has fewer than three indicators after adjusting the model, we usually keep only one question (indicator) for this factor and consider it as an observed variable as suggested by [91]. We choose Cronbach's alpha and correlated item-total correlations to measure the construct's internal reliability for considered latent variables. The scores of all constructs are above the acceptable level of 0.5 [92]. The scores of item-total correlations are above the cut-off value (0.4) for all constructs [92]. After several iterations, we obtained values as indicated in Table 2. They meet the cut-off values of all validity and reliability indicators. By running these validity tests, we refine each construct's questions and increase the validity of our evaluation model's constructs. After proving the model's reliability and validity, we test our hypothesized paths using the structural equation model (SEM) [93]. Table 2 shows thirty-seven question items that have been validated in our user study. By asking these question items, researchers can assess a CRS based on how users perceive both recommendations and conversations.

4.3 Structural Model

We employ the structural equation model (SEM) to build a path model for validating our hypothesized paths. Figure 2 shows the results of the structural model analysis. Overall, our model has a good fit indicated by the following indices, $\tilde{\chi}^2$ = 937.220 (d.f. = 580), $p < 0.001$, CFI = 0.923, TLI = 0.912, RMSEA = 0.060, which meet the recommended standard of these fit indices [94]. Besides, R squared values for most of constructs are larger than 0.30, which indicate that the model is able to examine significance of the paths associated with these constructs. To avoid the visual complexity of presenting all influencing paths, we keep the paths according to the hypothesized influence paths in ResQue (Perceived Qualities → User Beliefs → User Attitudes → Behaviral Intentions) and group some constructs of User Beliefs influenced by CUI Adaptability (a rectangle with dashed stroke). Since not all detected paths in SEM model are meaningful in terms of causal relationship [95], we also omit some paths that do not make sense for explaining the user experience of conversational recommender systems. Figure 2 shows all meaningful and significant paths in our model. The numbers on the arrows represent the β coefficients and standard errors of the effect. We distinguish recommendation constructs and conversation constructs using the white rectangle and the gray rectangle, respectively. The constructs with a red border are considered as the key qualities of a CRS because they tend to influence behavioral intention to use. Specifically, we identify two qualities of conversation (i.e., **Adaptability** and **Understanding**) and five qualities (i.e., **Accuracy**, **Novelty**, **Interaction Adequacy**, **Perceived Usefulness**, and **User control**) as key qualities of a CRS because they tend to influence behavioral intention to use.

For the key qualities of recommendations, the significant paths show that recommendation Accuracy, Novelty, and Interaction Adequacy positively influence Perceived Usefulness. For the key qualities of conversation, the significant paths indicate that the positive effects of CUI Adaptability on User Control and some non-key qualities (i.e., Perceived Ease of Use, Transparency, Rapport, and Engagingness); and the positive effects of CUI understanding on User control and CUI humanness. The significant paths (Perceived

Usefulness → Confidence, Perceived Usefulness → Overall Satisfaction, and Control → Trust) justify the positive effects of User Beliefs on User Attitudes. However, the conversation constructs do not influence the constructs in User Attitudes. The significant paths from Overall Satisfaction and Trust to Intention to Use indicate positive effects of User Attitudes on Behavioral Intentions. More notably, the SEM result indicate how key conversation qualities interact with recommendation qualities. For example, CUI Adaptability tend to positively influence several recommendation qualities (i.e., Perceived Ease of Use, Transparency, and Control), and CUI Understanding tends to positively influence User Control.

5 DISCUSSION AND LIMITATIONS

User-centric evaluation has gained extensive attentions in the recommender systems community. However, we found that conversational systems' evaluation work mainly considers objective metrics, such as understanding rate and dialogue turns. We argue that user perception of conversations strongly influences the overall user experience of recommendations. Therefore, we develop the model **CRS-UX** to capture key qualities of conversational recommender systems from the users' perspective. Our model seamlessly integrates several popular UX metrics of conversations into a widely used UX model of recommender systems [6] and reveals the synergy between recommendations and conversations in a CRS.

RQ1: Which qualities of a CRS are particularly important in terms of their influence on intention to use the CRS?

We first identify eight UX constructs for recommendations based on *ResQue* [6] and nine potential UX constructs for conversation based on existing research works. By performing confirmatory factor analysis (CFA), we merge several conversation constructs and integrate them into the two dimensions: User Perceived Qualities and User Beliefs. Ultimately, our **CRS-UX** model mainly accommodates two key conversation qualities (i.e., CUI Adaptability and CUI Understanding) and five key recommendation qualities (i.e., Accuracy, Novelty, Interaction Adequacy, Perceived Usefulness, and User Control). Besides, our model validates some previously verified paths, i.e., the positive effect of CUI attentiveness on CUI rapport [45] and the positive effects of accuracy, novelty, and interaction adequacy on perceived usefulness [6]. Our model also shows some additional paths between conversation constructs that suggest several promising ways to enhance humanness and engagingness, e.g.,increasing attentiveness in conversation, which might also be important to high user-involvement recommenders (for example, e-commerce recommenders) [96] that we will verify in the future studies.

RQ2: How do conversation constructs interact with recommendation constructs to influence user experience of CRSs?

The SEM results help us identify some new paths between recommendation constructs, which may imply the effects of conversational interaction on the relations among UX constructs of recommendations in *ResQue* model. For instance, we find that user control positively influences user trust, and overall satisfaction positively influences intention to use. Although conversation constructs do not directly influence the constructs of User Attitudes and Behavioral Intention, we find CUI Adaptability and CUI Understanding positively influence Trust and Intention to Use through the mediator User Control. Compared with a traditional recommender system, a CRS provides a more natural and free way to control the system, which, in turn, may increase user trust. Besides, we argue that the enjoyment of using conversational interaction may attract satisfied users to use the CRS repeatedly in the future [97]. More importantly, the paths among the constructs of conversations and recommendations explicitly show the added value of **CRS-UX** in explaining the user experience of conversational recommender systems. Interestingly, we find more explanations tend to increase rapport in conversations. This effect may imply an interesting research topic about designing conversational explanations for recommendations. Thus, these results validate a user-centric approach to determine key qualities of a CRS and suggest the possible ways to improve user experience of a CRS from both content (recommendations) and interaction (conversations) aspects.

Our work has several limitations: first, we validate **CRS-UX** with only one CRS for music. To examine the generalizability of our model, we may need to confirm the model's validity, reliability, paths, and structural consistency in a different application domain, for example, product recommendations. Besides, although the sample size in our study meets the minimum requirement (N=100–150) for conducting SEM [89], the larger sample size will increase the power of our study, considering the relatively large number of constructs in our model.

6 CONCLUSION

We propose a model **CRS-UX** to capture two key conversation qualities (i.e., Adaptability and Understanding) and five key recommendation qualities (i.e., Accuracy, Novelty, Interaction Adequacy, Perceived Usefulness, and User control) for conversational recommender systems. We review the subjective metrics of measuring UX of recommender systems and conversational agents and seamlessly integrate them into the four dimensions: Perceived System Qualities, User Beliefs, User Attitudes, and Behavioral Intentions. Thereby, our model can help practitioners determine the key qualities crucial to the overall user experience of conversational recommender systems (CRSs). We conduct an online user study to identify the validity and reliability of the constructs in our model. We adjust the model based on factor analysis results. Moreover, we identify several influencing paths that show how conversation constructs and recommendation constructs influence each other. We believe the questionnaire associated with our model can assess the usability of conversational recommender systems for different application domains. Eventually, we keep thirty-seven questions, which require moderate user effort to finish the questionnaire. As intelligent voice assistants become prevalent, such as Siri and Alexa, we think the key qualities of CRSs may also be constructive for designing and evaluating speech-based conversational recommenders regardless of the interaction modality. In the future, we plan to validate our model in other application domains (like e-commerce of high user involvement).

ACKNOWLEDGMENTS

The work was supported by two research grants: HKBU IRCMS/19-20/D05 and RGC/HKBU12201620.

REFERENCES

[1] Y. Jin et al. Musicbot: Evaluating critiquing-based music recommenders with conversational interaction. In *Proc. of CIKM'19*, pp. 951–960, 2019.

[2] W. Cai et al. Critiquing for music exploration in conversational recommender systems. In *Proc. of IUI'21*, pp. 480–490, 2021.

[3] S. M. McNee et al. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Proc. of CHI'06 EA*, pp. 1097–1101, 2006.

[4] J. A. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *UMUAI*, 22(1):101–123, 2012.

[5] M. Ge et al. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proc. of RecSys'10*, pp. 257–260, 2010.

[6] P. Pu et al. A user-centric evaluation framework for recommender systems. In *Proc. of RecSys'11*, pp. 157–164, 2011.

[7] B. P. Knijnenburg et al. Explaining the user experience of recommender systems. *UMUAI*, 22(4):441–504, 2012.

[8] F. Pecune et al. A model of social explanations for a conversational movie recommendation system. In *Proc. of HAI'19*, pp. 135–143, 2019.

[9] D. Jannach et al. A survey on conversational recommender systems. *arXiv preprint arXiv:2004.00646*, 2020.

[10] F. Ricci et al. *Recommender Systems Handbook*. Springer-Verlag, 2nd edition, 2015.

[11] K. Christakopoulou et al. Towards conversational recommender systems. In *Proc. of KDD'16*, pp. 815–824, 2016.

[12] L. Chen and P. Pu. Critiquing-based recommenders: Survey and emerging trends. *UMUAI*, 22(1-2):125–150, 2012.

[13] J. Kang et al. Understanding how people use natural language to ask for recommendations. In *Proc. of RecSys'17*, pp. 229–237, 2017.

[14] W. Cai and L. Chen. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *Proc. of UMAP'20*, pp. 33–42. ACM, 2020.

[15] Y. Sun and Y. Zhang. Conversational recommender system. In *Proc. of SIGIR'18*, pp. 235–244, 2018.

[16] K. Papineni et al. Bleu: a method for automatic evaluation of machine translation. In *Proc of ACL'02*, pp. 311–318, 2002.

[17] S. Fazeli et al. User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg. *IEEE Transactions on Learning Technologies*, 11(3):294–306, 2017.

[18] A. Said et al. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proc. of CSCW'13*, pp. 1399–1408, 2013.

[19] Y. Jin et al. Go with the flow: effects of transparency and user control on targeted advertising using flow charts. In *Proc. of AVI'16*, pp. 68–75, 2016.

[20] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *UMUAI, 22*(4-5):399–439, 2012.

[21] P. Pu and L. Chen. Trust building with explanation interfaces. In *Proc. of IUI'06*, pp. 93–100, 2006.

[22] B. P. Knijnenburg et al. Inspectability and control in social recommenders. In *Proc. of RecSys'12*, pp. 43–50, 2012.

[23] D. Braun et al. Evaluating natural language understanding services for conversational question answering systems. In *Proc. of SIGdial'17*, pp. 174–185, 2017.

[24] R. Dale and C. Mellish. Towards evaluation in natural language generation. In *Proc. of LREC'98*, 1998.

[25] A. Ghandeharioun et al. Approximating interactive human evaluation with self-play for open-domain dialog systems. 32:13658–13669, 2019.

[26] M. Walker et al. Paradise: A framework for evaluating spoken dialogue agents. In *Proc. of ACL'97*, pp. 271–280, 1997.

[27] M. Turunen et al. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In *Proc. of ICSLP'06*, 2006.

[28] Z. Ruttkay et al. Embodied conversational agents on a common ground. In *From brows to trust*, pp. 27–66. Springer, 2004.

[29] K. Kuligowska. Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, 2, 2015.

[30] N. Radziwill and M. Benton. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional*, 19(3):25, 2017.

[31] M. Guerini et al. A methodology for evaluating interaction strategies of task-oriented conversational agents. In *Proc. of EMNLP Workshop SCAI'18*, pp. 24–32, 2018.

[32] R. Zhao et al. Sogo: a social intelligent negotiation dialogue system. In *Proc. of IVA'18*, pp. 239–246, 2018.

[33] E. Svikhuushina and P. Pu. Key qualities of conversational chatbots – the peace model. In *Proc. of IUI'21*, pp. 520–530, 2021.

[34] M. Walker et al. Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3 & 4):363–377, 2000.

[35] P. Cremonesi et al. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *INTERACT'11*, pp. 152–168. Springer, 2011.

[36] N. Tintarev. Explanations of recommendations. In *Proc. of RecSys'07*, pp. 203–206, 2007.

[37] N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pp. 479–510. Springer, 2011.

[38] F. Gedikli et al. How should i explain? a comparison of different explanation types for recommender systems. *IJHCS*, 72(4):367–382, 2014.

[39] H. Cramer et al. The effects of transparency on trust in and acceptance of a content-based art recommender. *UMUAI*, 18(5):455, 2008.

[40] J. L. Herlocker et al. Evaluating collaborative filtering recommender systems. *ACM TOIS*, 22(1):5–53, 2004.

[41] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM TiiS*, 7(1):1–42, 2016.

[42] F. Narducci et al. Improving the user experience with a conversational recommender system. In *Proc. of AIxIA'18*, pp. 528–538. Springer, 2018.

[43] B. Priyogi. Preference elicitation strategy for conversational recommender system. In *Proc. of WSDM'19*, pp. 824–825, 2019.

[44] L. Chen and P. Pu. Interaction design guidelines on critiquing-based recommender systems. *UMUAI*, 19(3):167, 2009.

[45] L. Tickle-Degnen and R. Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.

[46] A. Abdellatif et al. A comparison of natural language understanding platforms for chatbots in software engineering. *arXiv preprint arXiv:2012.02640*, 2020.

[47] P. Kataria et al. User adaptive chatbot for mitigating depression. *International Journal of Pure and Applied Mathematics*, 118(16):349–361, 2018.

[48] J. Seering et al. It takes a village: Integrating an adaptive chatbot into an online gaming community. In *Proc. of CHI'20*, pp. 1–13, 2020.

[49] D. Wang and H. Fang. Length adaptive regularization for retrieval-based chatbot models. In *Proc. of SIGIR'20*, pp. 113–120, 2020.

[50] J. Jiang and N. Ahuja. Response quality in human-chatbot collaborative systems. In *Proc. of SIGIR'20*, pp. 1545–1548, 2020.

[51] A. See et al. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL-HLT (1)*, 2019.

[52] Y. Jin et al. Contextplay: Evaluating user control for context-aware music recommendation. In *Proc. of UMAP'19*, pp. 294–302, 2019.

[53] Y. Hijikata et al. A study of user intervention and user satisfaction in recommender systems. *Journal of information processing*, 22(4):669–678, 2014.

[54] D. Jannach et al. User control in recommender systems: Overview and interaction challenges. In *Proc. of EC-Web'16*, pp. 21–33. Springer, 2016.

[55] B. P. Knijnenburg et al. Each to his own: how different users call for different interaction methods in recommender systems. In *Proc. of RecSys'11*, pp. 141–148, 2011.

[56] Y. Jin et al. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *UMUAI*, 30(2):199–249, 2020.

[57] D. Mican et al. Perceived usefulness: A silver bullet to assure user data availability for online recommendation systems. *Decision Support Systems*, 139:113420, 2020.

[58] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pp. 139–183. Elsevier, 1988.

[59] I. Simonson. Determinants of customers' responses to customized offers: Conceptual framework and research propositions. *Journal of marketing*, 69(1):32–45, 2005.

[60] K. Swearingen and R. Sinha. Interaction design for recommender systems. In *Proc. of DIS'02*, volume 6, pp. 312–334. Citeseer, 2002.

[61] R. F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proc. of CHI'16*, pp. 2390–2395, 2016.

[62] R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *Proc. of CHI'02 EA*, pp. 830–831, 2002.

[63] D. Novick and I. Gris. Building rapport between human and eca: A pilot study. In *Proc. of HCI International'14*, pp. 472–480. Springer, 2014.

[64] L. D. Riek et al. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *JMUI*, 3(1):99–108, 2010.

[65] M. Lalmas et al. Measuring user engagement. *Synthesis lectures on information concepts, retrieval, and services*, 6(4):1–132, 2014.

[66] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *JASIST*, 59(6):938–955, 2008.

[67] D. Adiwardana et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

[68] E. Go and S. S. Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 2019.

[69] D. Westerman et al. I believe in a thing called bot: Perceptions of the humanness of "chatbots". *Communication Studies*, 70(3):295–312, 2019.

[70] H. Candello et al. Typefaces and the perception of humanness in natural language chatbots. In *Proc. of CHI'17*, pp. 3476–3487, 2017.

[71] R. M. Schuetzler et al. The impact of chatbot conversational skill on engagement and perceived humanness. *JMIS*, 37(3):875–900, 2020.

[72] N. Svenningsson and M. Faraon. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proc. of AICCC'19*,

pp. 151–161, 2019.
[73] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proc. of IUI'05*, pp. 167–174, 2005.
[74] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proc. of RecSys'07*, pp. 17–24, 2007.
[75] J. Kunkel et al. Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In *Proc. of CHI'19*, pp. 1–12, 2019.
[76] L. Chen and P. Pu. A cross-cultural user evaluation of product recommender interfaces. In *Proc. of RecSys'08*, pp. 75–82, 2008.
[77] L. Wang et al. When in rome: the role of culture & context in adherence to robot recommendations. In *Proc. of HRI'10*, pp. 359–366. IEEE, 2010.
[78] S. Berkovsky et al. How to recommend? user trust factors in movie recommender systems. In *Proc. of CHI'17*, pp. 287–300, 2017.
[79] E. J. De Visser et al. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331, 2016.
[80] A. Przegalinska et al. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6):785–797, 2019.
[81] J. Edwards and E. Sanoubari. A need for trust in conversational interface research. In *Proc. of CUI'19*, pp. 1–3, 2019.
[82] J. A. Hoxmeier et al. The impact of gender and experience on user confidence in electronic mail. *JOEUC*, 12(4):11–20, 2000.
[83] S. Z. Arshad et al. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of OzCHI'15*, pp. 352–360, 2015.
[84] T. T. Nguyen et al. User personality and user satisfaction with recommender systems. *Information Systems Frontiers*, 20(6):1173–1189, 2018.
[85] B. P. Knijnenburg and M. C. Willemsen. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proc. of RecSys'09*, pp. 381–384, 2009.
[86] L. Chen et al. How serendipity improves user satisfaction with recommendations? a large-scale user evaluation. In *Proc. of WWW'19*, pp. 240–250, 2019.
[87] Y.-Y. Wang et al. Understanding the moderating roles of types of recommender systems and products on customer behavioral intention to use recommender systems. *ISeB*, 13(4):769–799, 2015.
[88] D. Shin. How do users interact with algorithm recommender systems? the interaction of users, algorithms, and performance. *Computers in Human Behavior*, 109:106344, 2020.
[89] J. Wang and X. Wang. *Structural equation modeling: Applications using Mplus: Chapter 7.1.* John Wiley & Sons, 2019.
[90] T. A. Brown. *Confirmatory factor analysis for applied research: 3. Introduction to CFA.* Guilford publications, 2015.
[91] M. C. Willemsen et al. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *UMUAI*, 26(4):347–389, 2016.
[92] R. A. Peterson. A meta-analysis of cronbach's coefficient alpha. *Journal of consumer research*, 21(2):381–391, 1994.
[93] R. B. Kline and D. A. Santor. Principles & practice of structural equation modelling. *Canadian Psychology*, 40(4):381, 1999.
[94] D. Hooper et al. Structural equation modelling: Guidelines for determining model fit. *EJBRM*, 6(1):53–60, 2008.
[95] J. Pearl. The causal foundations of structural equation modeling. Technical report, UCLA DEPT OF COMPUTER SCIENCE, 2012.
[96] L. Qiu and I. Benbasat. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *JMIS*, 25(4):145–182, 2009.
[97] M. Heerink et al. Enjoyment intention to use and actual use of a conversational robot by elderly people. In *Proc. of HRI'08*, pp. 113–120, 2008.